

**SPEAKER IDENTIFICATION EMPLOYING A CONFIDENCE MEASURE**  
**THAT USES STATISTICAL PROPERTIES OF N-BEST LISTS**

**Field of the Invention**

The present invention generally relates to speaker identification systems,  
5 particularly, in which the speech of a given individual is analyzed and the identity of the  
individual is determined.

**Background of the Invention**

Speaker identification systems have been developed for years, and efforts continue  
to be made at improving upon prior versions. Several publications which provide but a  
10 small representation of the current state of the art include: D. A. Reynolds,  
“Experimental Evaluation of Features for Robust Speaker Identification”, IEEE  
Transactions on Speech and Audio Processing, Vol. 2, No. 4, pp. 639-643, 1994; D. A.  
Reynolds and R. C. Rose, “Robust Text-Independent Speaker Identification Using  
Gaussian Mixture Speaker Models”, IEEE Transactions on Speech and Audio Processing,  
15 Vol. 3, No. 1, pp. 72-83, January 1995; and U. V. Chaudhari, J. Navratil, S. H. Maes, and  
Ramesh Gopinath “Transformation Enhanced Multi-Grained Modeling for Text-  
Independent Speaker Recognition”, ICSLP 2000, pp. II.298-II.301.

Among the disadvantages observed in connection with conventional speaker identification systems is that such systems are generally not configured for being able to determine when a system result is inconclusive. Accordingly, a need has been recognized in connection with overcoming such disadvantages.

5 **Summary of the Invention**

In accordance with at least one presently preferred embodiment of the present invention, a speaker identification system is contemplated which is able to statistically model and evaluate whether a system result is inconclusive. In accordance with a preferred embodiment, an N-best list is analyzed and a confidence measure is obtained  
10 using statistical properties of the N-best list.

In summary, the present invention provides, in one aspect, an apparatus for facilitating speaker identification, said apparatus comprising: an arrangement for accepting input speech; an arrangement for generating at least one N-best list based on the input speech; an arrangement for positing a system output based on the input speech; and  
15 an arrangement for ascertaining, via at least one property of the N-best list, whether the posited system output is inconclusive.

Another aspect of the present invention provides a method of facilitating speaker identification, said method comprising the steps of: accepting input speech; generating at least one N-best list based on the input speech; positing a system output based on the input speech; and ascertaining, via at least one property of the N-best list, whether the  
5    posited system output is inconclusive.

Furthermore, the present invention provides, in an additional aspect, a program storage device readable by machine, tangibly embodying a program of instructions executable by the machine to perform method steps for facilitating speaker identification, said method comprising the steps of: accepting input speech; generating at least one N-  
10    best list based on the input speech; positing a system output based on the input speech; and ascertaining, via at least one property of the N-best list, whether the posited system output is inconclusive.

For a better understanding of the present invention, together with other and further features and advantages thereof, reference is made to the following description, taken in  
15    conjunction with the accompanying drawings, and the scope of the invention will be pointed out in the appended claims.

### **Brief Description of the Drawings**

Fig. 1 schematically illustrates a system of confidence-based speaker identification.

Fig. 2 schematically illustrates the generation of an N-best list in the context of

5 Fig. 1.

Fig. 3 schematically illustrates N-best list likelihood evaluation in the context of

Fig. 1.

### **Description of the Preferred Embodiments**

Throughout the present disclosure, various terms are utilized that are generally  
10 well-known to those of ordinary skill in the art. For a more in-depth definition of such  
terms, any of several sources may be relied upon, including Reynolds, Reynolds et al.,  
and Chaudhari et al., all *supra*.

Fig. 1 schematically illustrates a system of confidence-based speaker  
identification in accordance with an embodiment of the present invention. Input speech  
15 (102) is input into the speaker identification system 104. An N-best list 106 is then  
preferably generated, and sorted so that the first candidate is the one associated with the  
best score and the Nth candidate is the one with the N<sup>th</sup> lowest score, i.e. the worst score

among the top N candidates. In general, there will be a large population of enrolled speakers with size  $N_p \gg N$  and a score will be generated for all  $N_p$  speakers. The N-best list contains the N top scoring speakers (candidates). An objective will be to examine these lists and determine the level of confidence the system has as to the correctness of the best scoring candidate (108). Based on this measure, as queried at 110, either an answer is given (112) or the trial is determined to be inconclusive (114/116). If the trial is determined to be inconclusive, a determination is made as to whether the speaker is inconclusive on the whole (114) or if a further trial ("repeat trial") with more input speech from the same speaker is warranted (116). When more than one identification system is used, each analyzes the speech and the answer of the system with the highest confidence is used. (In some instances, it is desirable to use more than one identification system to avoid being limited to the particular type or range of scores generated, the statistical methods used on the N-best lists are not dependent on such parameters.)

With reference to Fig. 2, in accordance with at least one preferred embodiment of the present invention, for each system, two types of statistical models of each N-best list are generated. This modeling is separate from the modeling that is done of the acoustic properties in the speech signal. In fact, such acoustic models are preferably used in order to generate the scores (at 118) in the N-best lists.

Once the aforementioned two N-best list models are present, it is possible to evaluate the likelihood of the observed N-best lists with respect to both and incorporate the results in a procedure to evaluate the confidence in the top candidate (108 in Fig. 1). Reference may now be made to Fig. 3, which illustrates an N-best list being split into two  
5 models (124/126) inherent in a confidence scoring arrangement such as that indicated in Fig. 1 at 108.

Mathematically:

- Let  $s_1, s_2, \dots, s_N$  (indicated at 106a) be the top N scores ( $s_1$  is the best score).
- Let  $\mathbf{s} = \{s_1 - s_2, s_2 - s_3, \dots, s_{N-1} - s_N\}$  (the set of differences, preferably as generated  
10 via a difference generator 120).
- Let  $\mathbf{i} = i_1, i_2, \dots, i_N$  (indicated at 106b) be the N identities (arranged in order from best to worst).

Training is preferably accomplished with development data in the form of the candidate and score lists of a large set of trails (i.e., examples of 106a/b during real usage  
15 of the system, or "development data") where the lists are each split into two sets according to whether the top candidate is correct or incorrect. Note that this partition depends on the output of the identification system. Thus, one will be learning the

properties of the (acoustic) system output. The two N-best list models are preferably generated as discussed herebelow.

First, the set of development score difference vectors corresponding to the correct trials are preferably denoted  $\{s\}_{\text{correct}}$  and that of the incorrect trials is  $\{s\}_{\text{incorrect}}$ .

5       Next, one may preferably:

- construct a statistical model of  $\{s\}_{\text{correct}} = \mathbf{M}_{\text{correct}}$  ; and
- construct a statistical model of  $\{s\}_{\text{incorrect}} = \mathbf{M}_{\text{incorrect}}$  .

One can model each class (correct, incorrect) for example, as a Gaussian Mixture Model (GMM) (see Reynolds et al., *supra*); this is just one of many possibilities. In this  
10    case, the likelihood ratio would be used for the scoring of an observed N-best list, namely, the ratio of the likelihood with respect to  $\mathbf{M}_{\text{correct}}$  and likelihood with respect to  $\mathbf{M}_{\text{incorrect}}$ :

$$\text{likelihood ratio} = p(\{s | \mathbf{M}_{\text{correct}}\}) / p(\{s | \mathbf{M}_{\text{incorrect}}\})$$

where  $p(\{s | \mathbf{M})$  are the Gaussian densities. This ratio (and generation thereof) is  
15    schematically indicated at 124 in Fig. 3.

Next, with the second model, one will preferably estimate the likelihood of the actual identities in the N-best list. That is, for each enrolled speaker the objective is to be able to evaluate the likelihood of any given sequence of identities in the N-best list for a test trial with the speaker's data. Thus, using the training data for each speaker, one will  
5 construct a model of the composition of  $\{i\}$  correct for each speaker as follows:

Given a target  $m_i$ , for every model  $m$  in the enrolled population, estimate the probability that  $m$  is in the N-best list of a trial for which  $m_i$  is the correct answer. (This depends on N relative to the size of the total population and is a function of the average position of  $m$  in the ordered list  
10 of candidates for the training trials.)

For each development trial of each speaker, consider the  $N_p$ -best list (i.e. the ordered list of all of the identities and scores). Each enrolled model  $m$  has a position in this list. Preferably, the average position over all of the development trials for a given speaker will be computed. This average position can be interpreted as a "distance" to the  
15 top position. The position distribution of  $m$  is then preferably modeled with a Gaussian with mean and variance given by the average position and deviation from the development trials. Thus, for each pair of enrolled speakers, there will be a probability density for the position of one speaker in the other's N-best list. For testing, one may



assume independence and use the product model to evaluate the probability of the N-best list identities conditional on the top candidate being correct; thus:

$$likelihood \approx \prod_{i=2}^N p(i_j | i_1)$$

where  $p(i_j | i_1)$  is the aforementioned Gaussian density. (This quantity, and generation thereof, is schematically indicated at 126 in Fig. 3.)

Then, the two scores from (1) and (2) are preferably fused using Linear Discriminant Analysis (LDA), GMM, or a neural network model. (A good discussion of LDA and neural network models may be found in Pattern Classification and Scene Analysis, Duda and Hart, John Wiley & Sons, Inc. 1973.) This final score is the confidence measure. For concreteness, a linear combination (i.e. LDA) is preferably used (indicated at 128) to yield a confidence score (indicated at 130) as follows:

$$confidence = \alpha [likelihood\ ratio(1)] + \beta [likelihood\ ratio(2)]$$

This *confidence* can be compared to a threshold  $t_0$ , chosen so that a value less than  $t_0$  means that the system output for the trial should be considered inconclusive. In the broader framework, the speaker identification system may opt to collect more data from the individual and reevaluate the identities (step 116 in Fig. 1).



modifications may be affected therein by one skilled in the art without departing from the scope or spirit of the invention.

2013-03-26 10:35:00